STAT 401 Chapter 2.7, 2.8, 5.12, 6.5, 6.8(g), 7.1–7.4

Zepu Zhang
November 5, 2010

(The terminology and notation of this material is not identical to the textbook. You should follow this note.)

The intercept β_0 is the coefficient of a constant predictor 1. It is easy to understand that there is no need to take any other value for the constant predictor, and there can be at most one constant predictor in a model.

The textbook treats the constant predictor and the intercept as something rather special. In this note, we will try not to treat them so special. The constant predictor is a predictor just like any other predictor, and the intercept is just the parameter (coefficient) associated with this predictor.

Nevertheless, we'll keep the special notation: we may write the constant predictor as X_0 and the intercept as β_0. (Most likely we'll spell out "constant predictor" just to be clear.) The other predictors are $X_1, \ldots, X_{p-1}$, and coefficients $\beta_1, \ldots, \beta_{p-1}$.

Notation p is the number of β's (or predictor terms, or columns in the design matrix $\boldsymbol{X}$). It counts the intercept, if the model includes an intercept. n is always the number of observations.

With an "empty" model, our fitting for any Y is simply 0, which is not very useful. By introducing predictors, we attempt to describe/capture the variation in the observed Y. Note

$$Y_i = \hat{Y}_i + (Y_i - \hat{Y}_i)$$

or

$$\text{observation} = \text{fit} + \text{residual}$$

We would like to see much of the variation in Y being captured by $\hat{Y}$, and the residual being small. This will be how we measure the quality/significance of a model, or equivalently, the specific choice of predictors (assuming we have a number of candidate predictors).

We use sum of squares (SS) as an indicator of the amount of "variation" in a variable. Define

$$\begin{aligned} \text{SST (SS total):} &\quad \sum Y_i^2 \\ \text{SSR (SS due to regression, or model):} &\quad \sum \hat{Y}_i^2 \\ \text{SSE (SS due to error, or residual):} &\quad \sum e_i^2 \end{aligned}$$

What follows concentrates on how much of SST is captured in SSR. We check this by comparing SSR and SSE. If the former is large compared to the latter, the model is "good". We will make this assessment quantitative.

1 Sums of squares

We see

$$\text{SST} = \vec{Y}'\vec{Y}$$
$$\text{SSR} = \hat{\vec{Y}}'\hat{\vec{Y}} = (\boldsymbol{H}\vec{Y})'(\boldsymbol{H}\vec{Y}) = \vec{Y}'\boldsymbol{H}\vec{Y} \;\; (\text{which } = \hat{\vec{\beta}}'\boldsymbol{X}'\vec{Y})$$
$$\text{SSE} = (\vec{Y} - \hat{\vec{Y}})'(\vec{Y} - \hat{\vec{Y}}) = \vec{Y}'(\boldsymbol{I} - \boldsymbol{H})\vec{Y} = \vec{Y}'\vec{Y} - \vec{Y}'\boldsymbol{H}\vec{Y}$$

leading to the important relation

$$\boxed{\text{SST} = \text{SSR} + \text{SSE}}$$

SST has n degrees of freedom (df).

SSR has p df.

SSE has $n - p$ df.

Example If we use the model that has X_0 only, then we will have $\hat{\beta}_0 = \overline{Y}$. (Verify this!) Hence $\hat{Y} = \hat{\beta}_0 = \overline{Y}$.

Introduce notation $\boldsymbol{J} = \vec{1}\,\vec{1}'$ (a $n \times n$ square matrix of 1's), hence the n-column of $\overline{Y}$'s is $\frac{1}{n}\boldsymbol{J}\vec{Y}$. Then

$$\text{SSE} = (\vec{Y} - \tfrac{1}{n}\boldsymbol{J}\vec{Y})'(\vec{Y} - \tfrac{1}{n}\boldsymbol{J}\vec{Y}) = \vec{Y}'(\boldsymbol{I} - \tfrac{1}{n}\boldsymbol{J})'(\boldsymbol{I} - \tfrac{1}{n}\boldsymbol{J})\vec{Y} = \vec{Y}'(\boldsymbol{I} - \tfrac{1}{n}\boldsymbol{J})\vec{Y}$$

where we have used the fact $(\frac{1}{n}\boldsymbol{J})(\frac{1}{n}\boldsymbol{J}) = \frac{1}{n}\boldsymbol{J}$.

This SSE is the SYY defined earlier (i.e. $\sum(Y_i - \overline{Y})^2$). It is the sum of squares of Y about its mean. Hence it is also called SS corrected for the mean. We may denote it by $\text{SST}_{\overline{Y}}$ if needed.

Note that the design matrix $\boldsymbol{X}$ is now a column matrix with all 1's. It can be seen that $\boldsymbol{H} = \frac{1}{n}\boldsymbol{J}$, also leading to the SSE above.

This SSE is the residual sum of squares after using the "pure intercept model". In other words, this is the part of the variation in the observed Y that can not be captured by β_0 alone. Since the textbook on this topic assumes an intercept is always included in the model (and the issue is what other X's to pull in), this SS is what is left to be captured by the model. The textbook calls this term the SSTO.

2 Mean squares

Dividing each "sum of squares" by the associated "degrees of freedom" gives "mean squares":

$$\text{MSR} = \frac{\text{SSR}}{p}$$
$$\text{MSE} = \frac{\text{SSE}}{n-p}$$

We usually use the symbol s^2 for MSE. ("MST" is not very useful in this topic.)

To assess the significance of the regression model, instead of comparing SSR and SSE, we compare MSR and MSE, because we know the following about MSR and MSE.

Theorem

1. $E(\text{MSE}) = \sigma^2$. (i.e. MSE is an unbiased estimator of σ^2.)

2. MSR and MSE are independent of each other.

3. $E(\text{MSR}) \geq E(\text{MSE})$. (An exact formula for $E(\text{MSR})$ is known.)

4. $E(\text{MSR}) = E(\text{MSE})$ if and only if $\vec{\beta} = \vec{0}$. Note: $\vec{\beta} = \vec{0}$ means the true model is an empty one.

5. If $\vec{\beta} = \vec{0}$, then

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}}{p} \Big/ \frac{\text{SSE}}{n-p} \sim F_{\text{df(MSR)},\,\text{df(MSE)}}$$

that is, F has a F distribution with p, $n-p$ df. The two df's are the two parameters for the F distribution; the order of these two df matters.

About the F distribution: positive variable, two parameters,...

Now we can test the hypothesis $H_0 : \vec{\beta} = \vec{0}$. as follows.

Assume H_0 is true. Take

$$F = \text{MSR}/\text{MSE}$$

as a test statistic. If F is very large (extremely large in its $F_{p,n-p}$ distribution), then it suggests this F is not from the distribution $F_{p,\,n-p}$, hence $\vec{\beta} = \vec{0}$ is not true. Note an extremely large F suggests it's most likely true that $E(\text{MSR}) > E(\text{MSE})$, consistent with items 3–4 in the theorem above.

This is called a F test. We use the critical value

$$F(1-\alpha;\, p, n-p),$$

defined as the value such that the tail to its right (one-sided!) has area $1-\alpha$.

R tip The F critical value is calculated by `qf(1 - alpha, p, n - p)`.

Notice that the $\vec{\beta}$ tested here includes the intercept. It tests whether the model is useful (significant) at all. Often this test will reject H_0 because, even if the other X's are not very useful in predicting Y', the intercept is likely significant, unless the overall magnitude of Y does appear to be 0.

What would be more useful is to test whether the non-constant predictors are useful provided that the constant predictor (β_0) is already included in the model. Further, it would be nice if we could test whether a specific X brings significant additional contribution to the regression given that other predictors are already in the model, or whether certain predictors as a group make significant contribution.

3 Extra sums of squares

We will use the set of predictors to identify a model. For example, $\{X_0\}$ is a "pure intercept" model, whereas $\{X_1, X_2\}$ is the model with predictors X_1 and X_2, without intercept.

Suppose the full model is $\{X_0, X_1, \ldots, X_k\}$. Partition the predictors into two sets M_1 and M_2. ("Partition" means the two sets combine to give the full set, and they share no common elements.)

If we use the two sets of predictors separately, we have

$$\begin{aligned}\text{SST} &= \text{SSR}(M_1) + \text{SSE}(M_1)\\ \text{SST} &= \text{SSR}(M_2) + \text{SSE}(M_2)\end{aligned}$$

What happens if we first use the predictors in M_1, then add the predictors in M_2?

Theorem If we "grow" a linear model (meaning keep all existing predictors, add additional predictors), the SSR will always increase, and (necessarily) the SSE will always decrease.

We want to know what additional contribution the M_2 predictors make as we grow the model M_1 to $\{M_1, M_2\}$. Is the additional contribution "significant" so that it justifies pulling in the predictors M_2?

Since

$$\text{SST} = \text{SSR}(M_1, M_2) + \text{SSE}(M_1, M_2)$$

we see

$$\text{SSR}(M_1, M_2) - \text{SSR}(M_1) = \text{SSE}(M_1) - \text{SSE}(M_1, M_2)$$

Interpretation: the addition of M_2 results in an increase of the regression SS by the amount of $\text{SSR}(M_1, M_2) - \text{SSR}(M_1)$ and an decrease of the residual SS by the same amount (necessarily).

Definition We use the notation

$$\text{SSR}(M_2 \mid M_1) = \text{SSR}(M_1, M_2) - \text{SSR}(M_1) = \text{SSE}(M_1) - \text{SSE}(M_1, M_2)$$

and call it the extra sum of squares. It measures the contribution of the set of predictors M_2 to the regression SS when these predictors are added to the model that already contains predictors M_1.

The df of $\text{SSR}(M_2 \mid M_1)$ is the number of predictors in M_2.

Look again at the relations

$$\begin{aligned} \text{SSR}(M_1, M_2) &= \text{SSR}(M_1) + \text{SSR}(M_2 \mid M_1) \\ \text{SSE}(M_1) &= \text{SSE}(M_1, M_2) + \text{SSR}(M_2 \mid M_1) \end{aligned}$$

In words, informally,
(1) the regression SS due to $\{M_1, M_2\}$ is that due to M_1 alone plus the extra due to M_2 in addition to M_1.
(2) the error SS left by $\{M_1\}$ is reduced to the error SS left by the "fuller" model $\{M_1, M_2\}$, the amount of reduction being equal to the (extra) contribution of $\{M_2\}$ to the regression SS.

Example In SLR, We have
$\text{SSE}(X_0) = \text{SSE}(X_0, X_1) + \text{SSR}(X_1 \mid X_0)$ and
$\text{SSE}(X_1) = \text{SSE}(X_0, X_1) + \text{SSR}(X_0 \mid X_1)$. Here, $\text{SSE}(X_0) = \text{SST}_{\overline{Y}} = \text{SYY}$.

Example With predictors X_0, X_1, X_2, and X_3, we have relations such as
$\text{SSR}(X_2 \mid X_0, X_1) = \text{SSE}(X_0, X_1) - \text{SSE}(X_0, X_1, X_2)$,
$\text{SSR}(X_2, X_3 \mid X_0, X_1) = \text{SSE}(X_0, X_1) - \text{SSE}(X_0, X_1, X_2, X_3)$.

Example Suppose M_1, M_2, and M_3 are mutually exclusive groups of predictors, then

$$\text{SSR}(M_2, M_3 \mid M_1) = \text{SSR}(M_2 \mid M_1) + \text{SSR}(M_3 \mid M_1, M_2),$$

that is, the regress SS is added to stepwise. This can be seen from

$$\begin{aligned}\mathrm{SSR}(M_2, M_3 \mid M_1) &= \mathrm{SSE}(M_1) - \mathrm{SSE}(M_1, M_2, M_3) \\ &= \mathrm{SSR}(M_2 \mid M_1) + \mathrm{SSE}(M_1, M_2) - \mathrm{SSE}(M_1, M_2, M_3) \\ &= \mathrm{SSR}(M_2 \mid M_1) + \mathrm{SSR}(M_3 \mid M_1, M_2)\end{aligned}$$

Example Suppose $p = 3$, then
$\mathrm{SSR}(X1, X_2 \mid X_0) = \mathrm{SSR}(X_1 \mid X_0) + \mathrm{SSR}(X_2 \mid X_0, X_1)$.

4 F tests for β's based on sums of squares

Suppose M_1 is associated with coefficients $\vec{\beta}_1$ of length p_1, and M_2 with $\vec{\beta}_2$ of length p_2.

Theorem If $\vec{\beta}_2 = \vec{0}$, then

$$F = \frac{\mathrm{SSR}(M_2 \mid M_1)}{p_2} \Big/ \frac{\mathrm{SSE}(M_1, M_2)}{n - p_1 - p_2} \sim F_{p_2,\, n-p_1-p_2}$$

Since $\mathrm{SSR}(M_2 \mid M_1)$ is not obtained directly, but rather calculated as $\mathrm{SSR}(M_1, M_2) - \mathrm{SSR}(M_1)$, we may directly write

$$F = \frac{\mathrm{SSR}(M_1, M_2) - \mathrm{SSR}(M_1)}{p_2} \Big/ \frac{\mathrm{SSE}(M_1, M_2)}{n - p_1 - p_2} \sim F_{p_2,\, n-p_1-p_2}$$

Note the df of the numerator is $(p_1 + p_2) - p_1$, i.e. the number of "new" predictors (being tested on).

Equivalently,

$$F = \frac{\mathrm{SSE}(M_1) - \mathrm{SSE}(M_1, M_2)}{p_2} \Big/ \frac{\mathrm{SSE}(M_1, M_2)}{n - p_1 - p_2} \sim F_{p_2,\, n-p_1-p_2}$$

Calling $\{M_1\}$ the "reduced" model and $\{M_1, M_2\}$ the "full" model, we can say

$$F = \frac{\mathrm{SSE}_{\text{reduced}} - \mathrm{SSE}_{\text{full}}}{p_2} \Big/ \frac{\mathrm{SSE}_{\text{full}}}{n - p_1 - p_2}$$

The numerator concerns the reduction in the SSE due to the extra predictors, whereas the denominator concerns the SSE stilling remaining after introducing the extra predictors.

With this theorem, we can test
H_0: $\vec{\beta}_2 = \vec{0}$
vs
H_a: $\vec{\beta}_2 \neq \vec{0}$
in the context that the M_1 predictors are in the model.

Table 1: Testing $H_0 : \vec{\beta}_2 = \vec{0}$ for coefficients corresponding to predictor group M_2, given predictors M_1 already in model

Soure of variation	SS	df	MS (= SS/df)	F	crit. value
Regression	$\text{SSR}(M_2 \mid M_1)$	p_2	MSR	MSR/MSE	$\texttt{qf}(1-\alpha,\, p_2,\, n-p_1-p_2)$
Error	$\text{SSE}(M_1, M_2)$	$n-p_1-p_2$	MSE		
Total	$\text{SSE}(M_1)$	$n-p_1$			

* $\text{SSR}(M_2 \mid M_1) + \text{SSE}(M_1, M_2) = \text{SSE}(M_1)$.
* $p_2 + (n - p_1 - p_2) = n - p_1$.

Table 2: Testing $H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$, given intercept β_0 already in model

Soure of variation	SS	df	MS (= SS/df)	F	crit. value
Regression	$\text{SSR}(X_1, \ldots, X_{p-1} \mid X_0)$	$p-1$	MSR	MSR/MSE	$\texttt{qf}(1-\alpha,\, p-1,\, n-p)$
Error	$\text{SSE}(X_0, X_1, \ldots, X_{p-1})$	$n-p$	MSE		
Total	$\text{SSE}(X_0)$	$n-1$			

* $\text{SSR}(X_1, \ldots, X_{p-1} \mid X_0) + \text{SSE}(X_0, X_1, \ldots, X_{p-1}) = \text{SSE}(X_0)$.
* $(p-1) + (n-p) = n-1$.
* $\text{SSE}(X_0) = \text{SYY} = \sum_i (y_i - \overline{y})^2$, is the SSTO in the textbook, b/c $\hat{\beta}_0 = \overline{Y}$ and $\hat{y}_i = \overline{y}$ in the model $\{X_0\}$.

Let's put things in an ANOVA (ANalysis Of VAriance) table; see Table 1.

Note A vector $\neq \vec{0}$ means its elements are not all 0.

Example Let $M_1 = \{X_0\}$ and $M_2 = \{X_1, \ldots, X_{p-1}\}$, then the preceding test is about
H_0: $\beta_1 = \cdots = \beta_{p-1} = 0$.

In this case, the reduced model contains the intercept only. We have seen (earlier in this note) that

$$\text{SSE}_{\text{reduced}} = \vec{Y}'(\boldsymbol{I} - \boldsymbol{H})\vec{Y}$$

The textbook calls this SSTO. It is the residual SS left over after using the intercept. This is the starting point for all the tests in the textbook because the book assumes (for this topic) that the intercept is always included in the model.

Example Suppose the full model is $\{X_0\}$. Let $M_1 = \{\}$ and $M_2 = \{X_0\}$. Then $\text{SSE}_{\text{reduced}} = \vec{Y}'\vec{Y} = \text{SST}$ and $\text{SSE}_{\text{full}} = \vec{Y}'(\boldsymbol{I} - \boldsymbol{H})\vec{Y}$. Then $\text{SSR}(\{X_0\}) = \text{SSE}_{\text{reduced}} - \text{SSE}_{\text{full}} = \vec{Y}'\boldsymbol{H}\vec{Y}$, which is the contribution of the intercept alone. The mean-corrected SS takes $\vec{Y}'\boldsymbol{H}\vec{Y}$ out of $\vec{Y}'\vec{Y}$, hence it is "corrected for the mean or equivalently for the pure intercept model".

Remarks 1. We choose not to assume the intercept is always in the

model. One advantage is that, when we discuss grouping the predictors, a particular group may or may not include X_0. It's not different from the treatment of the other predictors.

2. If X_0 is used, then the residuals sum to 0. Subsequently, the mean of the fitted values equals the mean of the observations, i.e. $\sum Y_i = \sum \hat{Y}_i$. If X_0 is not used, we can not make these statements.

5 General linear test

The test above can be stated in more general forms. First, we have a "full" model:

$$Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 \cdots + \beta_{p-1} X_{p-1} + \epsilon$$

The model is "full" in two senses:

(1) This model contains all predictors we are going to consider in a particular question, and the question is "can we reduce this model?", i.e. "can we drop some of the X's (or β's)?"

(2) This full model is a valid linear model, meaning it satisfies the necessary model assumptions (normality, constant variance, linear relation, etc).

Suppose we want to test whether we can drop X_2. Then the assumption, H_0, is that in this full model the actual value of β_2 is 0, although in model fitting we'll get a non-zero estimate for β_2.

Assuming H_0 is true, then the reduced model

$$Y = \beta_0 X_0 + \beta_1 X_1 + \beta_3 X_3 + \cdots + \beta_{p-1} X_{p-1} + \epsilon$$

is still a valid linear model (because it is the same as the full model, noticing that $\beta_2 = 0$ in the full model). This validity as a linear model provides that the relations regarding sum of squares, and F tests, are valid.

(Some may say, "I have two other predictors, X_p and X_{p+1}, that are also related to Y and will make the model even better. Isn't $\{X_0, X_1, \ldots, X_{p+1}\}$ the real full model? The answer: the prospect that X_p and X_{p+1} would make a better linear model does not invalidate $\{X_0, X_1, \ldots, X_{p-1}\}$ as a linear model. We have a set of predictors that make a valid linear model, and our question is whether this model can be reduced. Our "full" model is full in this context.)

Given the reduced model above, our test statistic is

$$F = \frac{\text{SSE}_{\text{reduced}} - \text{SSE}_{\text{full}}}{\#\{\beta\}_{\text{full}} - \#\{\beta\}_{\text{reduced}}} \Big/ \frac{\text{SSE}_{\text{full}}}{n - \#\{\beta\}_{\text{full}}}$$

In the same vein, we can test more general forms of linear relations about the coefficients.

Example H_0: $\beta_2 = 2.5$.

Reduced model is

$$Y = \beta_0 X_0 + \beta_1 X_1 + 2.5 X_2 + \beta_3 X_3 \cdots$$

i.e.

$$Y - 2.5 X_2 = \beta_0 X_0 + \beta_1 X_1 + \beta_3 X_3 + \cdots$$

Example H_0: $\beta_2 = 2.5$ and $\beta_4 = 3$.

Similarly. The reduced model has 2 fewer coefficients than the full model.

Example H_0: $\beta_1 + \beta_2 = 2\beta_3$.

Reduced model is

$$Y = \beta_0 X_0 + (2\beta_3 - \beta_2) X_1 + \beta_2 X_2 + \beta_3 X_3 \cdots$$

i.e.

$$Y = \beta_0 X_0 + \beta_2 (X_2 - X_1) + \beta_3 (2X_1 + X_3) + \cdots$$

Or

$$Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \frac{\beta_1 + \beta_2}{2} X_3 + \ldots$$

i.e.

$$Y = \beta_0 X_0 + \beta_1 (X_1 + X_3/2) + \beta_2 (X_2 + X_3/2) + \beta_4 X_4 + \cdots$$

The two ways to represent the hypothesis are equivalent (both suggest the sum of the coefficients for X_1 and X_2 amounts to twice the coefficient for X_3). The reduced model has 1 fewer coefficient than the full model.

6 Comparison between t and F tests for $\vec{\beta}$

To recap, we can use a F test to test whether certain additional predictor(s) make significant contribution on top of an existing model. This hypothesis is often stated in the form of whether certain β's are 0. The existing model to begin with can be the totally null model ($E(Y) = 0$), can be the model with X_0 alone, can be a model with any set of predictors (including X_0 or not). The extra set of predictors to be tested on can be a single predictor, or any group of predictors. If it's a group of

extra predictors, the test is about the combined significance of the group as a whole.

We have learned t tests for the significance of any individual coefficient β_i, making use of the normal sampling distribution of β_i.

In comparison,

1. t test does a single β;
   F test does a single or a group of β's or linear relations between β's.

2. t test is against any hypothesized value (including and often 0);
   F test is against 0.

3. t test can be one-sided or two-sided;
   F test is two-sided (i.e. the alternative is always $\neq$).

The advantage of t test in item 2 is superficial: if we want to test $H_0 : \beta_k = 3$ for the coefficient of X_k, we can include a fixed term $3X_k$ in addition to X_k, and then test $H_0 : \beta_k = 0$.

The flexibility of t test in item 3 is not a big deal either—we usually know the sign of a coefficient from subject knowledge, all probably should leave it open if it's not clear-cut.

If we do a two-sided test against 0 on a single predictor, using a fixed α, the t and F tests are equivalent.

7 Coefficient of determination

The "coefficient of partial determination" with predictors M_2 given that predictors M_1 are already in the model is

$$R^2 = \frac{\mathrm{SSR}(M_2 \,|\, M_1)}{\mathrm{SSE}(M_1)}$$

Noticing the decomposition of $\mathrm{SSE}(M_1)$ into $\mathrm{SSE}(M_1, M_2) + \mathrm{SSR}(M_2 \,|\, M_1)$, we see $R^2 \leq 1$.

Following this definition, the R^2 with all the predictors $X_1, \ldots, X_{p-1}$ given that X_0 is in the model is

$$R^2 = \frac{\mathrm{SSE}(X_0) - \mathrm{SSE}(X_0, X_1, \ldots, X_{p-1})}{\mathrm{SSE}(X_0)} = 1 - \frac{\mathrm{SSE}(X_0, X_1, \ldots, X_{p-1})}{\mathrm{SSE}(X_0)}$$

where $\mathrm{SSE}(X_0) = \mathrm{SYY}$, as we know.

In SLR, R^2 is a useful indicator of how well the predictor X_1 models Y: the larger the R^2, the better the fitting.

In multiple regression, the second R^2 defined above will keep increasing if we introduce more and more predictors even if the marginal contribution of the additional predictors is small.

To cope with this, the "adjusted coefficient of determination" is defined as

$$R^2_{\text{adj}} = 1 - \frac{\text{SSE}(X_0, X_1, \ldots, X_{p-1})/(n-p)}{\text{SSE}(X_0)/(n-1)}$$

(Some modifications are needed if the model does not include intercept.) Now the numerator on the right-hand side does not necessarily increase if we introduce more predictors, because as SSE decreases, $n-p$ also gets smaller.

R^2_{adj} can be used as a crude model-selection tool. ("Model selection" here means choice of predictors.)

7.1 Coefficient of correlation

The square root of R^2 is called the "coefficient of partial correlation".

For univariate variables X and Y, define

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

(and we drop "partial" from the name).

Remarks

1. Rationale of this definition. Note $S_{xy} = \sum(x_i - \overline{x})(y_i - \overline{y})$: indicates whether X and Y tend to change concurrently in same or opposite directions.

2. Symmetric w.r.t. X and Y. It's a relation between X and Y. X and Y are in the same position, playing interchangeable roles.

3. $R^2 = r^2$.

4. $-1 \le r \le 1$.
   $r = 0$: no linear relation.
   $r \to 1$: strong positive linear relation.
   $r \to -1$: strong negative linear relation.
   $r = \pm 1$: points exactly on a straight line; exact linear relationship; one determines the other exactly.

Empirical, informal scales (nothing to be taken as a rule): $|r| \leq 0.5$: weak; $0.5 < |r| \leq 0.8$: moderate; $|r| > 0.8$: strong linear relationship.

7.2 Limitations of R^2 and r

Main caution about both r and R^2 is: both measure linear relationship only. Nonlinear relationship may well be present when $|r|$ and R^2 is small (or when they are large!). The best way to detect whether X and Y have nonlinear relationship is to look at a scatter plot.